Machine Learning with RandomForest: Baking a Serial Killer Victim Count Prediction Cake

##Clean data (remove NA values etc.). Then separate data into test and train set. Test and training sets are used to check for bias.

Remove NA Values

```
>SK5[is.na(SK5)] <- "0".
```

Turn character variables into factor variables

```
SK5 <- SK5 %>% mutate_if(is.character, as.factor)
```

Turn factors of numeric variables into numeric variables
First, make a list of all the columns you want to transform into numeric type

```
num <- c("Educ", "NumberofChildren", "Age1stKill","AgeLastKill", "NumVics")
```

##Next, use the sapply function to convert all columns in list to numeric type

```
SK5[num] <- sapply(SK5[num], numeric)
```

Cleaning is done. Split into test and training sets

```
> skVIC <- sample.split(SK5$NumVics, SplitRatio= 0.70)
> trainSKV <- subset(SK5, skVIC == T)
> testSKV <- subset(SK5, skVIC == F)
```

##Build Model with randomForest Package

```
rfSK1trV <- randomForest(NumVics ~Confessed + Sex + Race + US +
PreviousArrests + Previousjailorprisontime + KillwithHands + Weapon + VicSex +
RaceofVictim + VicAgeAdult + DadStable + MomStable + Married. +
Killerabusealcohol + Killerabusedrugs + BrainAb + HeadInj + Educ +
NumberofChildren + PsychAbuse + PhysAbuse + SexAbuse + Fired + Military +
Combat + BedWet + Rape + Torture + Soughtvictimtokill + Quick + Blindfold +
Bound + Mutilate + Totem + SexDeath + AteBody + DrankBlood + Posed +
BodyTotem + News + LeftHidden + LeftBuried + OrgDis, data=trainSKV)
```

#Make predictions on test set

```
predVIC <- predict(rfSK1trV, testSKV)
```

#Create dataframe to compare predicted values and test set values

```
realVIC <- as.data.frame(testSKV$NumVics)
```

```
compVIC <- cbind(realVIC,predVIC)

#To test for bias, one compares the MSE between the train and test sets for
predicted values. A spike in variance means the model is biased.

> trainPV <- as.data.frame(rfSK1trV$predicted)
> trainMSERF <- sqrt(sum(trainPV$`rfSK1trV$predicted`, -
trainSKV$NumVics)^2/nrow(trainSKV))

> trainMSERF
[1] 10.23006
#Not bad! That was the result for our train set. Let's compare against our test set.

> testMSERF <- sqrt(sum((predVIC$`predict(rfSK1trV, testSKV)` -
testSKV$NumVics)^2)/nrow(testSKV))
> testMSERF
[1] 17.6875

#The model is biased. The next step would be to compare against other models and
determine which model had the lest bias/variance.

#Gini Impurity

                    IncNodePurity
Confessed               34429.670
Sex                     17663.395
Race                    49511.670
US                      24788.226
PreviousArrests         14920.137
Previousjailorprisontime    13186.954
KillwithHands           25849.347
Weapon                  26348.491
VicSex                  56081.053
RaceofVictim            70047.155
VicAgeAdult             17421.883
DadStable               13946.273
MomStable               14423.124
Married.                17255.358
Killerabusealcohol      10957.587
Killerabusedrugs        12397.377
BrainAb                  4776.647
HeadInj                  4367.487
Educ                    28468.035
NumberofChildren        24924.975
PsychAbuse               7097.649
```

PhysAbuse	7127.497
SexAbuse	5260.298
Fired	3567.120
Military	12776.908
Combat	8415.916
BedWet	1611.811
Rape	20503.707
Torture	12778.283
Soughtvictimtokill	24568.848
Quick	21733.967
Blindfold	1697.929
Bound	15161.592
Mutilate	11143.832
Totem	7759.876
SexDeath	8656.233
AteBody	9362.116
DrankBlood	2077.915
Posed	5184.391
BodyTotem	3592.007
News	4693.447
LeftHidden	9906.023
LeftBuried	9119.095
OrgDis	36678.313